Data Immersion

Exercise 5.1

8/20/23

Micky Smith

Intro to Big Data

1. **Whats the difference between structured and unstructured data? Can you give examples that you've encountered for both types?**

   Structured data is organized data, often quantitative, containing rows and columns that can be presented by tables within databases. Unstructured data is disorganized and does not have a consistent format and can be difficult to analyze.
   Examples of structured data were provided in the Rockbuster database and Excel spreadsheets. It can include dates, addresses, credit card information, etc.
   Examples of unstructured data would be audio/video files, text, emails, etc.

2. **Given that much of big data is produced by machines and sensors, how trustworthy do you think that big data is? What characteristic of big data relates to the question of trustworthiness?**

   You can never be sure that the data provided is 100% reliable. Though the information is gathered by machine, the machines are created by humans and can still be subject to human error and bias. Because of that, it's important to be transparent in the method of collecting big data, or we cannot accurately assess it's trustworthiness.

3. **Assume that you receive a table containing the customer data. You notice that some values are missing or incomplete, and the formatting is inconsistent in some columns. Based on what you've learned so far, how would you go about cleaning this table? Think about what you would do first, second, third, etc.**

   I would start by checking each column for missing, incorrectly formatted, duplicated, or incomplete data. I would then reformat that data as needed and remove duplicates. I would then check for inaccurate data, and either delete or replace it depending on how the data needs to be handled for that data set. I would double check for missing data and confirm with the stakeholders how to proceed based on the company needs. And finish by making the data well rounded towards what would be meet the requirements of my research and how the stakeholders see the needs of the data set to work.

4. **Can you describe the tools such as Hadoop and Apache Spark and their role in big data? What do they do and how do they work?**

   Hadoop and Apache Spark are frameworks that can store and process big data and make it readily accessible for Data Analyst. It is done by distributing the data across a cluster of computers.

5. **How had the application of analytics to big data led to new discoveries and innovations? Can you give some examples?**

   Big data can be used to generate models which can be utilized to make predictions. Big data has the potential to be limitless as no ways of gathering and implementing data continues to grow. Big data can be used to predict weather patterns, severe weather, magnitude of catastrophic phenomena like hurricanes and earthquakes. Big data is also used to further the advancement of AI, big data gets fed to AI tools like ChatGPT in order to be provided tailored outputs for each response it provides.